DOCUMENT RESUME

ED 395 380 EA 027 569

AUTHOR Creemers, Bert P. M.; And Others
TITLE The International School Effectiveness Research Project Quantitative and Class Study Data, 1992-1994.
PUB DATE [96]
NOTE 27p.; For related paper see ED 376 206.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Academic Achievement; Analysis of Variance; *Effective Schools Research; Elementary Secondary Education; Ethnic Groups; Foreign Countries; Instructional Effectiveness; International Education; Mathematics Achievement; *Outcomes of Education; *School Effectiveness; Socioeconomic Status

ABSTRACT

This paper presents findings of a study, conducted by the International School Effectiveness Research Programme (ISERP), that examined school effectiveness in Australia, Canada, Hong Kong, Ireland, The Netherlands, Norway, Taiwan, the United Kingdom, and the United States. Data were analyzed at several levels: (1) at the individual level (student background, ability, and achievement); (2) at the classroom level (teacher behavior, curriculum, and class organization); (3) at the school level (organization, curriculum, and resources); and (4) at the contextual level. The study compared mathematics achievement as an outcome measure and, rather than looking at the differences among countries, looked at the differences among schools within countries. First-year results of the study indicate that students from Taiwan and Hong Kong mathematically outperformed students from other countries for which comparisons could be made (with the exception of Norway, where formal schooling starts 1 year later). Students' background variables, such as socioeconomic status, intelligence, and student ethnicity explained important proportions of student outcomes. In Taiwan and The Netherlands no significant differences existed after correction for intake; in other countries there were still substantial differences. In all the countries there was a relationship between observers' ratings of teachers and the classification of schools as effective or ineffective. Seven tables are included. (Contains 36 references.) (LMI)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 395 380

THE INTERNATIONAL SCHOOL EFFECTIVENESS RESEARCH PROJECT

QUANTITATIVE AND CLASS STUDY DATA

1992-1994

Bert P.M. Creemers
GION, Groningen Institute for Educational Research
University of Groningen
The Netherlands

David Reynolds
School of Education
University of Newcastle upon Tyne
United Kingdom

Frans E. Swint
GION, Groningen Institute for Educational Research

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Bert P.M. Creemers
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☐ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

The ISERP team is: Asbjorn Birkemo, Shin-Jen Chang, Yin Cheong Cheng, John Clarke, Bert Creemers, Barbara Dundas, Juanita Epp, Walter Epp, Barry Green, Marit Groterud, Vivian Hajnal, Trond Eiliv Hauge, Jen-jye Hwang, Peggy Kirby, Yong-yin Lee, Bjorn Nilsen, Hui-Ling Pan, David Reynolds, Sam Stringfield, Desmond Swan, Frans Swint, Charles Teddlie, Astrid Eggen Knutsen.

EA 027 569

BEST COPY AVAILABLE

1 Introduction

The history of educational effectiveness research can be described in different ways. A favourite way is to look at educational effectiveness as a reaction to the quite pessimistic views on teachers, schools and education in general brought forward by the disappointing results of research. Another, quite different interpretation of the history of educational effectiveness research considers this research as a natural prolongation of research from the past with respect to teaching, instruction, curriculum, school organisation, and so on. Depending on one's views of history different godfathers for educational effectiveness research are named, like Coleman et al. (1966) and Jencks et al. (1972). For most educational effectiveness research the work of Edmonds (1979) and Brookover, Beady, Flood, and Schweitzer (1979) in the United States and of Rutter, Maughan, Mortimore, and Ouston (1979) in the United Kingdom are important starting points for educational effectiveness research.

Especially in the United States, after Brookover and Edmonds a great deal of work has been done by researchers relating to the early American studies, but researchers in the United States seldom refer to any British or European literature other than that of Rutter et al. (1979) and Mortimore, Sammons, Stoll, Lewis, and Ecob (1988). In other countries researchers, although they refer to the earlier works of Brookover et al. (1979), Rutter et al., and Mortimore et al., show a similar ethnocentricity caused by the fact that in all these countries research was looking for factors that contribute to effectiveness within a specific country. The founding of the International Congress for School Effectiveness and Improvement and a journal associated with the organisation (*School Effectiveness and School Improvement*) in 1988, stimulated the exchange of research and research results in the field of educational effectiveness. The International School Effectiveness Research Programme (ISERP) is one of the examples.

2 National and international educational effectiveness studies

Out of the efforts of researchers in countries such as the United States, Great Britain, Canada, Norway, the Netherlands, Australia and New Zealand have come a large number of studies which look at the factors associated with student academic and social success within different countries, and within specific communities or 'ecological niches' within countries (see reviews in Bashi & Sass, 1992; Creemers, Peters, & Reynolds, 1989; Reynolds, Creemers, & Peters, 1989). When we have a look at the national studies that have been conducted first, it is clear that these have:

1 Tended to focus almost exclusively upon only the cognitive or 'academic' outcomes of schooling, and to have neglected therefore the affective or 'social' outcomes of the schooling process.

2 Tended to be focussed more upon gathering 'process data' at the level of the school rather than upon the level of the classroom, where instruction is centred.

3 Tended to be atheoretical, and to have been concerned with only the establishment of relationships between variables rather than with the generation and testing of theories which would account for, and explain, those relationships.

4. Tended to ignore any possible variation in those factors associated with student learning within different cultural contexts within countries, preferring instead the use of 'steampress' varieties of 'whole sample' analysis which aggregate and look at relationships across all schools (the United States research from Louisiana would be a notable exception to this - see Stringfield & Teddlie, 1990; and Wimpelberg, Teddlie, & Stringfield, 1989).

In addition to the defects in the national research bases that need to be remedied, the variation in the strength of the large number of national studies' research designs, research methodologies and in the mechanics of data collection in different societies makes the aggregation of the 'known to be valid' knowledge by which educational research progresses questionable, although recent effectiveness studies heavily depend on meta-analyses and reviews, including international studies that combine studies of different countries (Fuller & Clarke, 1994; Wang, Haertel, & Walberg, 1993). Within, and between, societies there are different conceptualisations of school and instructional process variables, different operationalisations even where the conceptualisation is similar and different measurement procedures and instruments even where the operationalisations are the same. Variations in sample characteristics, sampling techniques, methods of data collection, response rates, methods of analysis and statistical packages utilised increase the difficulty of the generation of any 'normal science' within the educational effectiveness research community, and reduce the possibility of any transfer of findings within and between countries at the level of educational policy. That much international educational policy discussion is currently simplistically based upon the transferability of factors shown to be effective in some cultures into the educational systems of others, merely increases the need to control the variation in the various national and international educational effectiveness research enterprises.

The findings of the nationally based educational effectiveness studies themselves also create a pressing need for further research to explore the apparent inconsistencies and complexities within the existing bodies of knowledge. Assertive principal instructional leadership, for example, is one of the most replicated of school process factors within the American research

literature in terms of being associated with positive student academic outcomes, yet within the Netherlands recent studies show no such principal effect (see review in Creemers, 1992). Out of the efforts of international organisations like the IEA, and out of surveys by the Educational Testing Service entitled the International Assessment of Educational Progress (IAEP), have also come a wide variety of studies that have looked at the between country variation in student outcomes from schools, and at the factors that may explain variations between countries (IEA examples are Anderson, Ryan, & Shapiro, 1989; Postlethwaite & Wiley, 1992; Robitaille & Garden, 1989; Travers & Westbury, 1989. IAEP examples are Keys & Forman, 1989; Lapointe, Mead, & Phillips, 1989.).

The studies carried out by the IEA and IAEP are especially focussed on student outcomes at a certain point in their educational career and on making these outcomes comparable between countries. They include information about student and teacher background variables, that can be analysed in relation to the outcomes. Although the international studies of the IEA also contain information about process factors, these are not the main focus of the studies. IEA studies are especially interesting in their continuous attention for and still improving measurement of 'opportunity to learn' as an indicator for the provision of education in specific countries. Other variables at the school and classroom levels, connected with the quality of education, however, cannot get the appropriate measurement in this kind of large-scale internationally comparative studies (see for a further, more elaborated, discussion of international evaluation studies: Reynolds, Creemers, Bird, & Farrell, 1994).

However, the studies contain information that is useful for the development of an international comparative set of indicators, especially with respect to outcomes and input and to some extent to process indicators like time and opportunity. The studies cannot provide information about the 'black box' educational process areas which have not been studied either in 'nationally' or 'internationally' based educational effectiveness research. The linkages between the levels of the classroom, or the instructional level, and the level of school processes have hardly been explored, yet it is precisely the study of this interface which the use of, and findings from, multilevel methodologies has rendered necessary (see for an exception: Bosker & Scheerens, 1994). Within-school variation by classroom in learning gains is now established to vary widely in some schools and less so in others - what 'interface' at the levels of the classroom and the school may be the explanation of this? This is an important question for research to pursue (See for a theoretical model of the relationship between levels: Creemers, 1994).

The educational factors outside the school at 'meso' or local community levels are also relatively unexplored by the 'national' and 'international' studies except for the interesting work of Coleman and Laroque (1990). What factors within the district or local education authority or 'local educational state' have effects upon schools and their classrooms, in terms

of potentiating or hindering pupil development? The effects of local economic, social and cultural meso context in terms of community influences, rather than the narrow educational context as above, also need investigation, particularly since some research shows considerable variation in what makes for 'effectiveness' in different ecological contexts (e.g. Wimpelberg, Teddlie, & Stringfield, 1989). There are also whole groups of school input variables (e.g. school financial and physical resources) which have been historically neglected (in this case because of the now heavily dated literature reviews from the 1980s which used aggregated measures of resources and pupil outcomes, except for the recently published meta-analysis by Hedges, Laine, & Greenwald (1994), which suggested a larger influence of resources on educational outcomes than the famous analysis by Hanushek (1989)) and which need further exploration.

3 The International School Effectiveness Research Programme (ISERP)

The ISERP programme of studies aims to do a number of things. It aims to build on existing models of 'good practice' in terms of research design, and aims to avoid the variation in national studies' research designs that limits transferability within and between countries, by utilisation of standard measures of intake, processes and outcomes, common methods of data analysis and common methods of data collection (although each participating country can add on to the common core any samples, instruments and analyses that may be appropriate for its own knowledge and/or policy needs).

A number of studies have been initiated as part of the overall programme. Firstly, the past three years have seen an extended programme of meetings, the production of discussion papers and of presentations to many academic conferences, as we have sought to make sense of the educational effectiveness literature, the international school/instructional effectiveness studies and the varying contexts, both socio-cultural and educational, of the different countries involved. Publications based on this phase have already appeared (Creemers, Reynolds, Schaffer, Stringfield, & Teddlie, 1991; Reynolds et al., 1994).

Secondly, there is the present pilot study which attempts, through the study of outlier schools in different national cultural contexts, to both develop and test hypotheses concerning the variables at different levels (class, school) which have an impact on student learning.

4 Research questions and design of the first study

Generally speaking, the research questions can be phrased as follows:

1 Which factors are associated with student academic and social outcomes within a particular cultural context within each particular society within each country (e.g. within low SES context).

2 Which factors are in general associated with student academic and social outcomes across countries and what factors are restricted to a certain cultural context (e.g. within low SES context across countries, etc.).

3 Which factors are associated with academic and social outcomes across countries by different student characteristics.

In the design of the research programme several models for educational effectiveness were compared. According to Ralph and Fennessey (1983), it was important not to find out just which factors make a difference between effective and ineffective schools, which leads to a fishing expedition for all kinds of variables (see for example Levine & Lezotte, 1990), but the development of models that make clear distinctions between context, input, process and outcomes and between the different levels in the educational system too, such as context, school, classroom and student. Based on the general model developed earlier by Scheerens and Creemers (1989), both authors developed models for school effectiveness (Scheerens, 1992) and educational effectiveness (Creemers, 1994) in which they try to combine instructional effectiveness at the classroom level and the conditions at the school level and the contextual level. A comparable model is the QAIT model by Stringfield and Slavin (1992) in which a distinction is made between quality, appropriateness, incentives and time. Although these models differ with respect to the number of variables and factors included, or with respect to the emphasis on the individual level, the classroom level or the school level, they have in common that they make a distinction between time and/or opportunity to learn and the instructional quality of schooling. These broad categories are elaborated at each of the levels within the educational system (see Figure 1).

In the educational effectiveness research programme within ISERP but also in several national studies these kinds of models are being tested.

The countries involved in this pilot study are Australia, Canada, Hong Kong, Ireland, the Netherlands, Norway, Taiwan, the United Kingdom, and the United States. The design of the study is mainly determined by a lack of funding which makes it impossible to include national representative samples of schools.

The study consists of a more quantitative and a more qualitative part. The qualitative part includes case studies within countries and a comparison of these case studies between

countries. The quantitative study has a design that could be used for a larger longitudinal survey study, but is now applied to a small sample of schools. In each country effective, averagely effective, the so-called 'typical' schools, and ineffective schools are involved in this study. Within each group of schools a distinction is made between low SES and middle SES schools. From each of these categories at least one school, including two classes, is involved. In some countries, like the Netherlands where schools normally do not have two classes per grade, just one class is included. Of course, this makes a comparison with respect to a distinction between class and school factors more difficult.

The selection of schools was based on an analysis of existing databases, when available. Otherwise, schools were selected on the basis of reputational criteria as identified by those key informers in the educational system who are likely to recognise positive, average and negative outliers. This can always be checked in the analysis of the data later on. The gainscore data will be used to validate the classification of the schools into one of the three effectiveness categories.

The data collection takes place at the individual student level with respect to student background, ability, and achievement in both academic and social areas; at the classroom level with respect to teacher behaviour, the curriculum and the organisation of the class; at the school level again with respect to the curriculum, the organisation of the school and the resources; and finally at the contextual level. The collection of data started in the schoolyear 1992/1993 and will be continued during the schoolyear 1993/1994, including research into the transition from one grade level to another. To make an international comparison not too difficult mathematics was chosen as an academic outcome measure. At the start the children were 7 years old, which implies that in most of the countries they are in the second grade of primary school where there is more emphasis on mathematics. Between countries, however, there are slight differences, such as in Norway where children start with primary education at the age of 7 and therefore do not have any mathematical background.

5 Results

5.1 Student outcomes

Differences between schools

The main point of interest is not the diffrence between various countries but the differences between schools within the various countries. These differences form the basis for a comparison of the effectiveness of schools. Based upon the design described above every

country sampled schools. Because of drop-out due to research overload the final samples in the countries varied between 5 and 12 schools, with 1 or 2 classes per school.

To compute the difference between the various schools the score on outcome testing year 2 is used. This score is considered to be the result of the teaching of mathematics over the years the students attended the schools. In all countries there were highly significant differences between the schools on these scores.

These differences can be accounted for by two categories of factors: the student background and the processes going on in the schools. Because the schools have different student intake it can be expected that this differential intake produces differences in outcome that are not caused by school processes. So, to compare schools it is necessary to correct for the intake of students. The factors considered relevant in this respect are the student's intelligence, the socioeconomic status (SES) of the family the student comes from and the question whether the student is a member of an ethnic minority group.

The intelligence of the students was tested with a 40 item non verbal intelligence test. The reliability (Cronbach's alpha) of this test is 0.83 based upon a sample of 2526 students. 582 students took the test two times with a year in between. The correlation between these two testing occasions was 0.68. For SES data were collected about the employment status of the parents, their jobs and their education. This was done for both the female and the male caretaker. Based upon these data the SES of the parent was coded in the categories low, mid and high. The correlation between the SES of the male caretaker and the female caretaker was 0.66. The students SES was defined as the average score of these two or as the score of the parent about whom information was available in the cases this was only one. Ethnicity was coded as a dummy variable with 0 meaning not a member of an ethnic group and 1 meaning a member of an ethnic group.

For every country a multivariate analysis of covariance was carried out to test the hypothesis that there were differences between schools when student background had been taken into account. The raw scores on outcome testing were used as criteria for both computation and applications. The variables described were used as covariates on the condition that they met the following criteria: at least a significant (one tailed, $p < .05$) correlation with one of the criteria and a relation in the hypothesized direction with the other criterium (positive for intelligence and SES and negative for ethnicity). The results of these analyses are summarized in Table 1. This table presents for every country the multivariae effect of the covariates, the multiple R for each of the two criteria, the test for multivariate differences between schools and a univariate test of differences between schools when the covariates were taken into account for each of the two criteria.

As can be concluded from Table 1 the covariates explain important proportions of the student outcome results. From the table it can be concluded that the variance between schools in the

sample in Taiwan and the Netherlands is mostly caused by differences in student intake. After correcting for this intake there are no significant multivariate differences between schools. This is also the case in Canada, but for this country the low power of the test has to be taken into account. After correction for covariates the other countries still show substantial differences between schools. It should be noticed that these effects are in all cases on both parts of the mathematics test, with only Norway being an exception.
The differences between schools can also be expressed as the proportion of the total variance explained by the factor school. This would mean adopting a multilevel approach in which the variance is divided between two levels: the level of the student and the level of the school. For each level the amount of variance (or the proportion) can be computed. Given the current research design, this would mean that these proportions show the relative contribution of each of the levels to the test results of the student. Because of the sampling procedures used, these proportions do not necessarily have to be the same as the proportions for the entire country. Tables 2 and 3 summarize the contribution of the student and school levels to the end of year 2 outcome testing results of the students. The variance components are given both as the estimated variance and as the percentages of the total variance. The components are computed twice: once without covariates and once with the covariates mentioned in Table 1 taken into account.

Attitudes towards school

Apart from data on academic outcomes of schooling, data were collected on the students' attitudes towards school and learning. A series of questions were given about the students' attitudes towards school, the teachers, mathematics and fellow students. All items were 5 point scales. Higher scale points were marked with smiling faces, lower scale points with sad faces. Also, questions were asked about the students' democratic attitudes. These items covered opinions about who made decisions in class and whether students were permitted to give their opinions. The items were given on a three point rating scale, again presented as smiling faces. A fourth series of items measured the locus of control of the students. These were items which described situations. For each situation, the student had to give an opinion on which of two options had most likely caused the situation. For each item, one option described an internal cause and one option described an external cause.

Factor analysis of the items about the student attitudes gave for both school years a three factor solution. The first factor can be interpreted as attitude towards school, the second factor as self-concept and the third as attitude towards mathematics. These factors were found by an analysis over all students. This three factor solution can be found by separate factor

analyses for each of the countries too. In case of exceptions, one finds factors which can be interpreted as social behaviour towards teachers, classmates or both. However, there are not many exceptions. It was decided to work further with the following scales: attitude towards school, self-concept, attitude towards mathematics, democratic attitudes and locus of control. The number of items and the reliabilities for these scales are given in Table 6. Also, the correlation between the first and second school year is given.

As can be seen in Table 6, the correlations between the attitude measurements in year 1 and the measurements in year 2 are rather low. Because the correlations between the scales were higher, it was decided to carry out a factor analysis over all attitude scales as collected in both school years. A principal component analysis resulted in 4 factors with an eigenvalue larger then 1. These factors were varimax rotated. The rotated solution is presented in Table 7.

Table 7 shows that the first component has primarily loadings from scales measured at the second year. The second component has scales measured at the first year. The third component covers the locus of control scales, the fourth the democratic attitude scales. It is unexpected to find components which are so heavily influenced by the time of measurement. Also, one should notice that the scales which load on separate factors (locus of control and democratic attitudes) both had different types of items than the other scales. So, an interpretation that the factors are influenced by method and time artifacts cannot be excluded. This makes it difficult to interpret further analyses with these scales. These results suggest that the students, who were 8 or 9 years old during test taking, do not yet have clearly formed attitudes.

School effects can be estimated by performing an analysis of variance for every scale with school as factor. Regarding the scales, many inconsistencies were found. For every scale, a significant school effect was found in about half of the countries. When one considers the individual countries, the most significant school effects are found in Great Britain, Norway (both 7 out of 8) and the USA (3 out of 4). In Hong Kong and Taiwan 5 significant effects were found out of 8 scales, in the Netherlands 3 out of 8. For these three countries, significant differences are not consistent over time. A difference found in one year might not mean that the schools also differ the next year.

The correlation between the student's attitude towards school and the computation score in both years is 0.10 (with a higher score on attitudes meaning a more positive attitude towards school). The correlation of the attitude towards school and the application test is 0.07 in both

years. For both years,the correlation between locus of control and each part of the mathematics test is about -0.27. Given the fact that a more positive score on the scale means a more external locus of control, this finding suggests that students who explain events more by internal causes, score slightly higher on mathematics tests. The correlation between self-concept and the mathematics test is almost zero. The correlation between mathematics attitudes and the scores on the mathematicas tests is about 0.14. The correlation between the democratic attitudes and computation score is about 0.10, the correlation between democratic attitudes and application is about 0.07. Given the fact that all these correlations were computed over the entire sample, all correlations except the self-concept ones, are significant. However, the values of the correlations are low. When one computes the correlations for every country separately, most coefficients decrease or approach zero.

5.2 Classroom Processes

As noticed in the previous section, there are some differences between the schools in the samples. An interesting question is whether these differences can be attributed to differences in classroom processes. Therefore, observations of lessons in mathematics were carried out. In the following the results of the observational studies in the first year of the project will be described.

One of the instruments used in the observation of the classroom was a rating scale. This scale was an adaptation of the Virgilio rating scale (Teddlie, Virgilio & Oescher, 1990). It contained 45 items, with each item describing teacher behaviour. The list of teacher behaviours described in the items was based upon previous research of effective teacher behaviour. Items covered the following domains: classroom management, maintaining order, student practice, questioning skills, teaching methods and classroom climate. During the project various versions of the instrument were used with either 3 point or 5 point scales. In all cases the highest scale point assumed that the behaviour was used frequently or excellent. For every class information was available from between 1 and 4 observations.

To analyze the instrument, the scores on all the different versions of the instrument were recoded to scores between 0 and 1 by dividing the score by the highest possible score on an item. In classes of which more observations were available, the average score for each item was computed over all observations. This was done to get a more stable picture of the classroom. In total, information was available on 93 classes in all countries. A principal components factor analysis was carried out to investigate the structure of the instrument. The eigenvalues and percentages of variance explained of the first 5 factors were 19.0 (42%), 2.9 (6.5%), 2.6 (5.8%), 2.4 (5.3%) and 2.2 (4.8%) with 5 other factors having an eigenvalue larger then 1 based upon the scree criterium. This suggests a one factor instrument. Rotations

to 3 and 5 factor solutions did not produce factors that could be clearly interpreted. The reliability of the 45 items scale (Cronbach's alpha) is 0.97. This also suggests a scale which is unidimensional.

The consequence of the one factor scale is clear: teachers either display the teaching behaviours originally described as effective or do not display these behaviours. On the basis of this material it cannot be concluded that there are teachers who differ in teaching styles, for example, being mostly focussed on the management of classes or mostly on interactive teaching. The data suggest an all or nothing differentiation of teachers. This is more noticable when one realizes that the teachers work in different countries which have different school systems. Unfortunately, the number of observations per country does not permit factor analysis per country.

5.3 Observations and results

Given the fact that there are some differences between the schools and differences in teacher behaviour, one can ask whether differences in teacher behaviour are related to differences in the performance of students. Therefore, for each teacher four scores are computed: a score for the original classification of effectiveness, two empirical effectiveness scores and one score for the teacher observations.

The original effectiveness classification of the teacher is assumed to be the same as the original effectiveness classification of the school the teacher works for. The effectiveness of the teacher is computed both for the computations and the applications part of the mathematics test. For each test a regression analysis was carried out with the score the student achieved on the mathematics test at the end of project year 1 as a criterium and as predictors the intake of that year and the covariates mentioned for that country in Table 1. The residual of this regression analysis is considered as the added value the teacher has above the expected outcome of the student. This added value is averaged for every student in the class to get a score for the teacher. The observation score is defined as the mean score of the teacher on all the items of the observation scale.

Since the number of classes in the various countries was rather small, it did not seem appropriate to conduct a multilevel analysis with predictors at the class level. Therefore, all classes were rank ordered on each of the variables described above. The relation between variables is given as the correlation between the rank orders of the classes (spearman's rho). The results for this analysis are given in Table 4.

Table 4 leads to some conclusions: first, in most countries the effectiveness computed for this school year is highly consistent with the effectiveness of the school as defined during the sampling of the schools. Second, the effectiveness of the school is in most cases related to

the observed teacher behaviour in the classes, for both the empirical and the original effectiveness classification.

Studying the observation scale one can ask whether there are differences in the quality of education as perceived by the observers in various countries. To compute these differences the mean score on the observation scale was computed for each country. These scores are presented in Table 5.

In general the mean scores for Taiwan, Norway and the Netherlands are about 0.10 higher than the scores for the USA, Great Britain and Hong Kong. This means a significant difference between countries ($F(5,87) = 7.94$, $p < .05$) in perceived teaching quality.

6 Conclusions

ISERP data are now analysed for outcomes in mathematics for the first year of the study of all the countries involved and partly for the process data: the results on the Virgilio rating scale for 6 countries. With respect to mathematics students from Taiwan and Hong Kong outperform students in all the other countries whose results are comparable (with the exception of Norway, but in that country formal schooling starts one year later).

Students' background variables, SES, intelligence and student ethnicity explain important proportions of student outcome results; in Taiwan and the Netherlands no significant differences exist after correction for intake; in other countries there are still substantial differences. These results are supported by an HLM analysis carried out to illustrate the proportion of variance.

The teaching processes in the classroom are rated by observers. This rating scale represents one factor: teaching quality. In all the countries there is a relationship between these ratings by observers and the classification of schools. A further elaboration based upon the items of the Virgilio reveals differences in the different components of teaching quality, which may be related to differences in educational (class, school, system) culture of the countries involved. However, further analysis of the available data is needed.

Acknowledgement

The authors would like to thank Dr. Hans Kuyper, GION, Groningen Institute for Educational Research, for his valuable contribution to the data analysis plan.

References

Anderson, L.W., Ryan, D.W.J., & Shapiro, B.J. (1989). *The IEA Classroom Environment Study*. Oxford: Pergamon Press.

Bashi, J., & Sass, Z. (Eds.). (1992). *School Effectiveness and Improvement: Proceedings of the Third International Congress for School Effectiveness, Jerusalem*. Jerusalem: The Magnes Press.

Bosker, R.J., & Scheerens, J. (1994). Alternative Models of School Effectiveness Put to the Test. *International Journal of Educational Research, 21*(2), 159-180.

Brookover, W.B., Beady, C., Flood, P., & Schweitzer, J. (1979). *School systems and student achievement: schools make a difference*. New York: Praeger.

Coleman, J.S., Campbell, E.Q., Hobson, C.F., McPartland, J., Mood, A.M., Weinfeld, F.D., & York, R.L. (1966). *Equality of Educational Opportunity*. Washington, DC: US Government Printing Office.

Coleman, P.J., & Laroque, L. (1990). *Struggling to be Good Enough*. Lewes: Falmer.

Creemers, B.P.M. (1992). School effectiveness and effective instruction: the need for a further relationship. In J. Bashi & Z. Sass (Eds.), *School Effectiveness and Improvement. Proceedings of the Third International Congress, Jerusalem* (pp. 105-132). Jerusalum: The Magnes Press.

Creemers, B.P.M. (1994). *The effective classroom*. London: Cassell.

Creemers, B.P.M., Peters, T., & Reynolds, D. (Eds.). (1989). *School Effectiveness and School Improvement. Proceedings of the Second International Congress, Rotterdam*. Lisse: Swets & Zeitlinger.

Creemers, B.P.M., Reynolds, D., Schaffer, E.C., Stringfield, S., & Teddlie, C. (1991). *International School Effects Research*. Kaohsiung: National Kaohsiung Normal University Press.

Edmonds, R.R. (1979). Effective schools for the urban poor. *Educational Leadership, 37*(1), 15-27.

Fuller, B., & Clarke, P. (1994). Raising school effects while ignoring culture? Local conditions and the influence of classroom tools, rules, and pedagogy. *Review of Educational Research, 64*, 119-157.

Hanushek, E.A. (1989). The impact of differential expenditures on school performance. *Educational Researcher, 18*(4), 45-65.

Hedges, L.V., Laine, R.D., & Greenwald, R. (1994). Does money matter? A meta-analysis of studies of the effects of differential school inputs on student outcomes (An exchange: Part 1). *Educational Researcher, 23*(3), 5-14.

Hill, P.W., Holmes-Smith, P., & Rowe, K.J. (1993). *School and Teacher Effectiveness in Victoria: Key Findings from Phase 1 of the Victorian Quality Schools Project*. Melbourne: Centre for Applied Educational Research, The University of Melbourne Institute of Education.

Jencks, C., Smith, M., Acland, H., Bane, M.J., Cohen, D., Gintis, H., Heyns, B., & Michelson, S. (1972). *Inequality: a reassessment of the effects of family and schooling in America*. New York: Basic Books.

Keys, W.J., & Forman, D. (1989). *A World of Differences: A U.K. Perspective on an International Assessment of Mathematics and Science*. Oxford: Pergamon Press.

Lapointe, A.E., Mead, N., & Phillips, G. (1989). *A World of Differences: An International Assessment of Mathematics and Science*. Princeton, N.J.: Educational Testing Service.

Levine D.U., & Lezotte, L.W. (1990). *Unusually effective schools: a review and analysis of research and practice*. Madison (USA): National Center for Effective Schools Research and Development.

Mortimore, P., Sammons, P., Stoll, L., Lewis, D., & Ecob, R. (1988). *School Matters: The Junior Years*. Wells: Open Books.

Postlethwaite, T.N., & Wiley, D.E. (1992). *The IEA Study of Science II: Science Achievement in Twenty-Three Countries*. Oxford: Pergamon Press.

Ralph, J.H., & Fennessey, J. (1983). Science or reform: some questions about the effective schools model. *Phi Delta Kappan, 64*(10), 689-694.

Reynolds, D., Creemers, B.P.M., Bird, J., & Farrell, S. (1994). School Effectiveness - The Need for an International Perspective. In D. Reynolds, B.P.M. Creemers, P.S. Nesselrodt, E.C. Schaffer, S. Stringfield, & C. Teddlie (Eds.), *Advances in School Effectiveness Research and Practice* (pp 183-201). Oxford: Pergamon Press.

Reynolds, D., Creemers, B.P.M., Nesselrodt, P.S., Schaffer, E.C., Stringfield, S., & Teddlie, C. (Eds.). (1994). *Advances in School Effectiveness Research and Practice*. Oxford: Pergamon Press.

Reynolds, D., Creemers, B., & Peters, T. (Eds.). (1989). *School Effectiveness and Improvement: Proceedings of the First International Congress, London, 1988*. Groningen: University of Groningen, RION.

Robitaille, D.F., & Garden, R.A. (1989). *The IEA Study of Mathematics II - Contexts and Outcomes of School Mathematics*. Oxford: Pergamon Press.

Rutter, M., Maughan, B., Mortimore, P., & Ouston, J. (1979). *Fifteen Thousand Hours*. London: Open Books.

Scheerens, J. (1992). *Effective Schooling: Research, Theory and Practice*. London: Cassell.

Scheerens, J., & Creemers, B.P.M. (1989). Conceptualizing school effectiveness. *International Journal of Educational Research*, 13, 691-706.

Scheerens, J., Nanninga, H.C.R., & Pelgrum, W.J. (1989). Generalizibility of instructional and school effectiveness indicators across nations; preliminary results of a secondary analysis of the IEA second mathematics study. In B.P.M. Creemers, T. Peters, & D. Reynolds (Eds.), *School Effectiveness and School Improvement. Proceedings of the Second International Congress, Rotterdam* (pp. 199-209). Lisse: Swets & Zeitlinger.

Stringfield, S.C., & Slavin, R.E. (1992). A hierarchical longitudinal model for elementary school effects. In B.P.M. Creemers & G.J. Reezigt (Eds.), *Evaluation of Educational Effectiveness* (pp. 35-69). Groningen: ICO.

Stringfield, S., & Teddlie, C. (1990). School improvement efforts: qualitative and quantitative data from four naturally occuring experiments in Phases 3 & 4 of the Louisiana School Effectiveness Study. *School Effectiveness and School Improvement, 1*, 139-161.

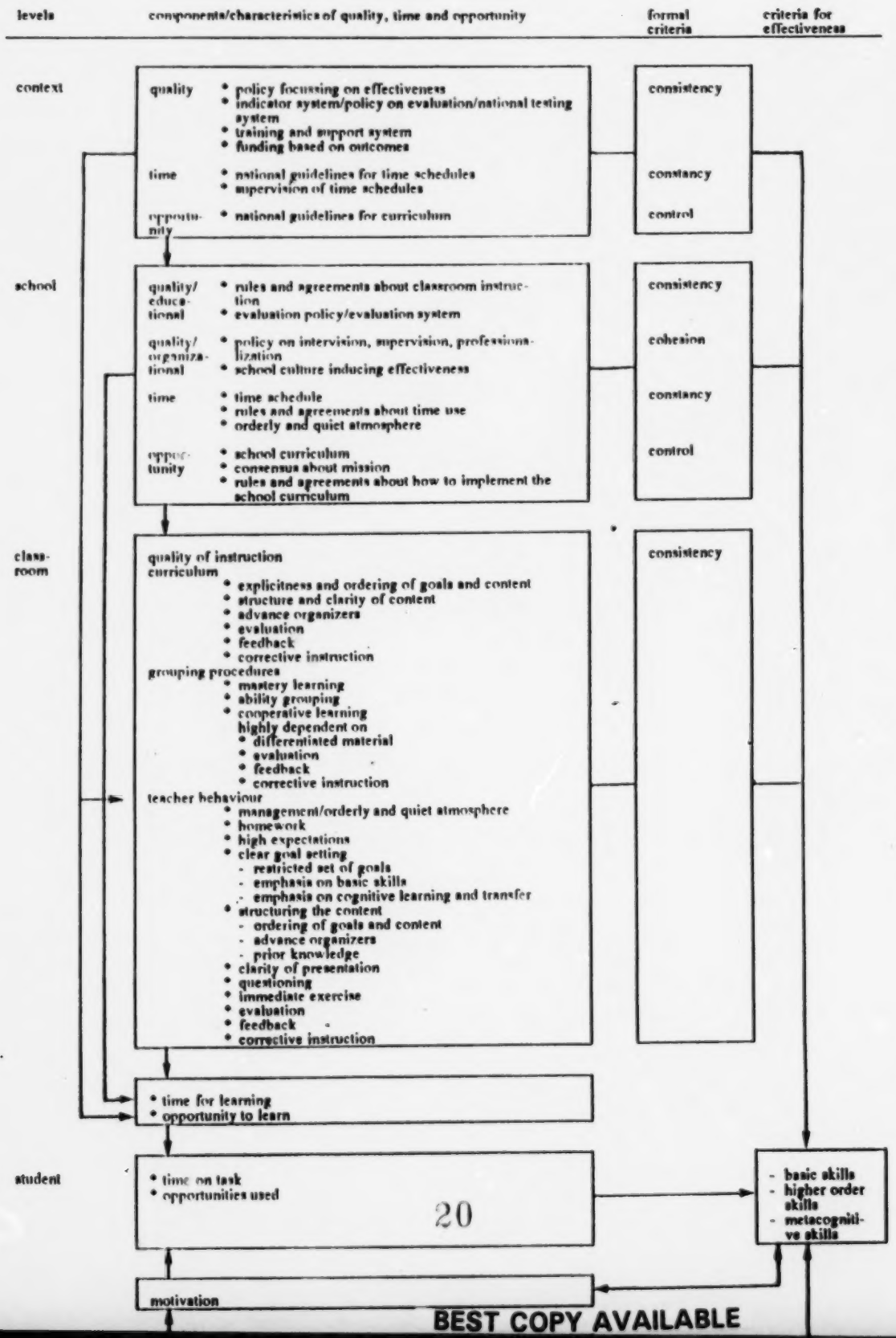
Teddlie, C., Virgilio, I., & Oescher, J. (1990). Development and Validation of the Virgilio Teacher Behavior Instrument. *Educational and psychological measurement, 50*, 421-430.

Travers, K.J., & Westbury, I. (1989). *The IEA Study of Mathematics I - Analysis of Mathematics Curricula*. Oxford: Pergamon Press.

Wang, M.C., Haertel, G.D., & Walberg, H.J. (1993). Toward a knowledge base for school learning. *Review of Educational research, 63*, 249-294.

Wimpelberg, R., Teddlie, C., & Stringfield, S. (1989). Sensitivity to context: the past and future of effective schools research. *Educational Administration Quarterly, 25*, 82-107.

Figure 1 A comprehensive model of educational effectiveness



BEST COPY AVAILABLE

Table 1: School Effects on Outcome Measures

Country	Covariates				School effects		
	Multivariate	Covariates	R Comp	R Appl	Multivariate	Computation	Applications
USA	F(6,950)= 18.3 p<.05	IQ SES ETHNIC	0.42	0.40	F(22,950)=8.4 p<.05	F(11,475) = 10.0 p<.05	F(11,475)= 5.9 p<.05
Great Britain	F(4,724)= 23.9 p<.05	IQ SES	0.43	0.43	F(22,724)=4.0 p<.05	F(11,362)= 5.0 p<.05	F(11,362)=4 .1 p<.05
Taiwan	F(4,824)= 29.8 p<.05	IQ SES	0.37	0.50	F(12,824)=1.4 N.S.	F(6,412)=2 .1 p=.05	F(6,412)=1. 7 N.S.
Norway	F(6,594)= 15.8 p<.05	IQ SES ETHNIC	0.43	0.45	F(16,594)= 2.0 p<.05	F(8,297)=1 .4 N.S.	F(8,297)=3. 3 p<.05
Hong Kong	F(4,524)= 14.5 p<.05	IQ SES	0.36	0.42	F(8,524)=8.0 p<.05	F(4,262)=8 .4 p<.05	F(4,262)=2. 9 p<.05
Canada	F(2,62)= 13.2 p<.05	IQ	0.52	0.47	F(8,126)=1.9 N.S.	F(4,63)= 1.4 N.S.	F(4,63) =3.9 p<.05
Nether-lands	F(4,226)= 7.4 p<.05	IQ ETHNIC	0.36	0.47	F(10,226)=1.3 N.S.	F(5,113)=2 .2 .05<p<.10	F(5,113)=1. 3 N.S.

Legend:
Section covariates:
Multivariate: test of multivariate significance of covariates
Covariates: Covariates used in the analysis for this country
R Comp: Multiple correlation coefficient for computation part of the test
R Appl: Multiple correlation coefficient for applications part of the test
Section School effects:
Multivariate: test of multivariate significance of difference between schools when covariats are taken into account
Computation: test of significant differences between schools on computation score corrected for covariates
Application: test of significant differences between schools on application scores corrected for covariates

Table 2: Variance Components Computation Test

COMPUTATION	No covariates		Controlled for covariates	
	Student level	School level	Student level	School level
USA	44.4 (82%)	9.9 (18%)	40.7 (83%)	8.6 (17%)
Great Britain	63.1 (81%)	14.7 (19%)	51.4 (89%)	5.9 (11%)
Taiwan	52.9 (98%)	1.1 (2%)	45.7 (99%)	0.6 (1%)
Norway	44.6 (99%)	0.3 (1%)	36.7 (99%)	0.2 (1%)
Hong Kong	62.2 (90%)	6.9 (10%)	54.1 (90%)	5.9 (10%)
Canada	46.3 (100%)	0 (0%)	35.5 (95%)	2.0 (5%)
Netherlands	33.5 (89%)	4.0 (11%)	29.1 (96%)	1.3 (4%)

Table 3: Variance Components Applications Test

APPLICATIONS	No covariates		Controlled for covariates	
	Student level	School level	Student level	School level
USA	66.7 (83%)	13.6 (17%)	61.6 (84%)	11.7 (16%)
Great Britain	52.1 (82%)	11.7 (18%)	42.6 (92%)	3.9 (8%)
Taiwan	73.6 (98%)	1.6 (2%)	55.1 (99%)	0.4 (1%)
Norway	34.5 (94%)	2.3 (6%)	27.6 (95%)	1.6 (5%)
Hong Kong	38.9 (99%)	0.3 (1%)	32.1 (98%)	0.8 (2%)
Canada	65.2 (95%)	3.2 (5%)	54.7 (89%)	6.5 (11%)
Netherlands	45.6 (90%)	5.3 (10%)	35.7(100%)	0 (0%)

Table 4: Relationship Between Effectiveness Criteria and Observation Measure

		EFFECT COMPUT	EFFECT APPLIC	EFFECT DEFINED
USA	Effect Application	.50 (42) *		
	Effect Defined	.45 (42) *	.16 (42)	
	Observation	.21 (15)	-.19 (15)	.50 (15)
Great Britain	Effect Application	.64 (23) *		
	Effect Defined	.67 (23) *	.42 (23) *	
	Observation	.35 (23) *	.54 (23) *	.45 (23) *
Norway	Effect Application	.58 (19) *		
	Effect Defined	0 (19)	.18 (19)	
	Observation	.34 (19)	.39 (19) *	.31 (19)
Hong Kong	Effect Application	.27 (12)		
	Effect Defined	.27 (12)	.09 (12)	
	Observation	.17 (12)	.36 (12)	-.35 (12)
Netherlands	Effect application	.48 (8)		
	Effect Defined	.08 (8)	.28 (8)	
	Observation	-.10 (8)	.64 (8) *	.47 (8)

effect computation: effectiveness score based on residual computation test
effect application: effectiveness score based on residual application test
effect Defined: original effectiveness classification
Observation: score on the observation ratings
All coefficients are spearman's rank order correlations computed over classes. Coefficients marked with * are significant ($p < .05$). The number in () is the number of classes used to compute the coefficient. Numbers smaller then in table 1 are caused by missing values in parts of the data set.

Table 5: Average observation rating per country

Country	Mean observation (classes)
USA	0.62 (16)
Great Britain	0.64 (24)
Taiwan	0.79 (14)
Norway	0.78 (19)
Hong Kong	0.67 (12)
Netherlands	0.77 (8)
Ireland	0.75 (7)

Table 6: characteristics of attitude scales

Scale	No of Scale points	No of items	Reliability	Test / Retest
Attitude towards school	5	6	.71	.36
Self-Concept	5	5	.66	.35
Mathematics attitude	5	2	.59	.38
Democratic Attitude	3	8	.61	.33
Locus of Control	2	10	.67	.48

Legend:
Scale = scale name, Reliability expressed as Cronbach's alpha, test/restest = correlation between testing year1 and year2

Table 7: rotated factor solution of attitude scales

	Factor 1	Factor 2	Factor 3	Factor 4
Att y1	.14	.73	-.21	.06
Att y2	.75	.21	-.10	.09
Loc y1	-.05	-.07	.83	-.16
Loc y2	-.02	-.08	.81	-.02
Self y1	.16	.74	.18	.19
Self y2	.77	.08	.13	.22
Math y1	.20	.69	-.14	-.09
Math y2	.72	.21	-.12	-.16
Democ y1	-.09	.32	-.03	.78
Democ y2	.25	-.17	-.20	.74

Legend: Att = attitude toward school, Locus = locus of control, self = self concept, math = attitude toward mathematics, democ = democratic attitude y1= measured in year 1, y2 = measured in year 2.